A finite algorithm for the switching control stochastic game

Vrieze, O.J.; Tijs, S.H.; Raghavan, T.E.S.; Filar, J.A.

Published in:
Operations Research Spektrum

Publication date:
1983

OR Spektrum (1983) 5:15–24

OR Spektrum

A Finite Algorithm for the Switching Control Stochastic Game

O. J. Vrieze[1], S. H. Tijs[1], T. E. S. Raghavan[2], and J. A. Filar[3]

[1] Department of Mathematics, Catholic University, Nijmegen, The Netherlands,
[2] Department of Mathematics, University of Illinois, Chicago, USA,
[3] Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, USA

Received 17, March 1982 / Accepted in revised from 26, October 1982

Summary. In this paper two-person zero-sum stochastic games are considered with the average payoff as criterion. It is assumed that in each state one of the players governs the transitions. We will establish an algorithm, which yields in a finite number of iterations the solution of the game i.e. the value of the game and optimal stationary strategies for both players. An essential part of our algorithm is formed by the linear programming problem which solves a one player control stochastic game. Furthermore, our algorithm provides a constructive proof of the existence of the value and of optimal stationary strategies for both players. In addition, the finiteness of our algorithm proves also the ordered field property of the switching control stochastic game.

Zusammenfassung. Wir betrachten stochastische Zweipersonen-Nullsummenspiele mit der durchschnittlichen Auszahlung als Kriterium. Wir nehmen an, daß in jedem Zustand einer der Spieler das Übergangsgesetz kontrolliert und entwickeln einen Algorithmus, der nach endlichen vielen Iterationsschritten die Lösung des Spiels – d. h. den Spielwert und optimale stationäre Strategien für beide Spieler – liefert. Ein wesentlicher Teil unseres Algorithmus besteht aus dem linearen Programm, das ein stochastisches Spiel löst, bei dem ein Spieler das Übergangsgesetz bestimmt. Darüber hinaus geben wir mit unserem Algorithmus einen konstruktiven Beweis der Existenz des Spielwertes und optimaler stationärer Strategien für beide Spieler. Weiter zeigt die Endlichkeit unseres Algorithmus die "ordered field property" stochastischer Spiele mit wechselnder Kontrolle des Übergangsgesetzes.

1. Introduction

In 1975 Parthasarathy and Raghavan began studying the class of two-person zero-sum stochastic games, where one of the players controls the transitions in all states. Their interest was in finding suitable algorithms for this class of games. Their first result was the fact that for discounted games of this type, there exists an LP-algorithm and that the value of such a game lies in the same ordered field as the other game parameters.

Stern [13] also in 1975 established the existence of the value for such games in the undiscounted case. This result was also obtained by Bewley and Kohlberg [1] in 1976.

In 1976 Parthasarathy and Raghavan proved, that for the undiscounted case the orderfield property also holds and they gave an algorithm for the irreducible case. They presented these results at the Dynamic Programming Conference in Vancouver in 1976. The results appeared recently in [11].

In 1979 Filar and Raghavan [6] found an algorithm for the undiscounted one player control games and presented their results at the Oberwolfach Conference on Game Theory in 1980. However, that algorithm was not very efficient.

An efficient LP-algorithm was found in 1980 independently by Vrieze [14] and Hordijk and Kallenberg [9], based on minimal harmonic functions.

In his Ph.D. dissertation, Filar [4] also proved that for discounted and undiscounted stochastic games, with switching controls, the ordered field property holds [5]. This indicated that for the switching control case, a finite algorithm should also exist and a first attempt to find such an algorithm was made in [7].

The purpose of this paper is to provide an efficient algorithm for the undiscounted game.

2. Definitions and Notations

A switching control stochastic game (notation Γ) is characterized by a seven-tuple

$$\langle S, S_1, S_2, \{A_{1k}; k \in S\}, \{A_{2k}; k \in S\}, r, p\rangle.$$

Here $S := \{1, 2, \ldots, N\}$ (the state space) and $A_{nk} := \{1, 2, \ldots, m_{nk}\}$ (the set of pure actions for player n in state k). S_1 and S_2 are subsets of S such that $S_1 \cap S_2 = \emptyset$ and $S_1 \cup S_2 = S$. (S_n is the subset of states were player $n \in \{1, 2\}$controls the transitions). r is a realvalued function (the payoff function) on the set $T := \{(k, i, j); k \in S, i \in A_{1k}, j \in A_{2k}\}$. [In this paper the variable i will always refer to a pure action of player 1 and j to a pure action of player 2.] The realvalued function $p: U^1 \cup U^2 \to \mathbb{R}$ (the transition law), where

$$U^1 := \{(l|k, i);\ l \in S,\ k \in S_1,\ i \in A_{1k}\}$$

$$U^2 := \{(l|k, j);\ l \in S,\ k \in S_2,\ j \in A_{2k}\}$$

has the properties

$p(l|k, i) \geqslant 0$ for all $k \in S_1$, $i \in A_{1k}$ and

$$\sum_{l \in S} p(l|k, i) = 1,$$

$p(l|k, j) \geqslant 0$ for all $k \in S_2$, $j \in A_{2k}$ and

$$\sum_{l \in S} p(l|k, j) = 1.$$

The interpretation of these parameters is as follows: if in state $k \in S$, the players take pure actions i and j respectively, then player 1 obtains an immediate payoff $r(k, i, j)$ from player 2 and the system moves with probability $p(l|k, i)$ to state $l \in S$ if $k \in S_1$ and with probability $p(l|k, j)$ if $k \in S_2$.

A strategy for player 1 (player 2) in the infinite stage game is denoted by π_1 (π_2). A stationary strategy for player 1 (player 2) is denoted by σ (ρ). Then $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$, where $\sigma_k = (\sigma_k(1), \sigma_k(2), \ldots, \sigma_k(m_{1k}))$ is a mixed action for player 1 in state k, choosing pure action $i \in A_{1k}$ with probability $\sigma_k(i)$.

By $\mathrm{car}(\sigma_k)$ we mean the set $\{i \in A_{1k};\ \sigma_k(i) > 0\}$. $\mathrm{Car}(\rho_k)$ has an analogous meaning.

For a pair of stationary strategies (σ, ρ) we introduce $N \times N$-matrices $P_{\sigma\rho}$ and $Q_{\sigma\rho}$ and the N-vector $r_{\sigma\rho}$. Here $P_{\sigma\rho}$ is the $N \times N$-transition matrix, with in the (k, l)-th entry the number

$$p(l|k, \sigma) := \sum_{i \in A_{1k}} p(l|k, i)\sigma_k(i) \quad \text{if } k \in S_1, \quad \text{and}$$

$$p(l|k, \rho) := \sum_{j \in A_{2k}} p(l|k, j)\rho_k(j) \quad \text{if } k \in S_2.$$

$Q_{\sigma\rho}$ is the Cesaro-limit of $P_{\sigma\rho}$, i.e.

$$Q_{\sigma\rho} := \lim_{t \to \infty} (t+1)^{-1} \sum_{n=0}^{t} P_{\sigma\rho}^n,$$

where $P_{\sigma\rho}^0$ is the identity matrix and $P_{\sigma\rho}^n = P_{\sigma\rho} \cdot P_{\sigma\rho}^{n-1}$ for $n \geqslant 1$. $r_{\sigma\rho}$ is the vector in $\mathbb{R}^N$, with k-th coordinate

$$r(k, \sigma_k, \rho_k) := \sum_{i \in A_{1k}} \sum_{j \in A_{2k}} r(k, i, j)\sigma_k(i)\rho_k(j).$$

The meaning of $r(k, \sigma_k, j)$ and $r(k, i, \rho_k)$ will be obvious.

Note that the following property holds:

$$Q_{\sigma\rho} = P_{\sigma\rho} \cdot Q_{\sigma\rho} = Q_{\sigma\rho} \cdot P_{\sigma\rho}.$$

[For detailed information we refer to Derman [2].]

$V(\pi_1, \pi_2)$ will stand for the limit expected average payoff for a pair of strategies (π_1, π_2). $V(\pi_1, \pi_2)$ is an N-vector, where the k-th coordinate corresponds to the particular stochastic game with state k as starting state. Note that if π_1 and π_2 are both stationary strategies, say σ and ρ, then $V(\sigma, \rho) = Q_{\sigma\rho} r_{\sigma\rho}$.

The game is said to have a value, if coordinatewise

$$\sup_{\pi_1} \inf_{\pi_2} V(\pi_1, \pi_2) = \inf_{\pi_2} \sup_{\pi_1} V(\pi_1, \pi_2).$$

A strategy π_1^* (π_2^*) for player 1 (player 2) is called optimal if the value (say V) exists and if

$$\inf_{\pi_2} V(\pi_1^*, \pi_2) \geqslant V \quad (\sup_{\pi_1} V(\pi_1, \pi_2^*) \leqslant V).$$

Further notation, which will be used is $[f(i, j)]_{B \times C}$, which denotes a matrix game with entries $f(i, j)$, $(i, j) \in B \times C$. The value such a game will be denoted by $\mathrm{Val}_{B \times C}[f(i, j)]$.

By an extreme optimal action for player $h \in \{1, 2\}$ in a matrix game, we mean an extreme point of the convex set of optimal mixed actions of player h. From Shapley and Snow [12], we know that the set of optimal mixed actions of a player in a matrix game is a polytope, so there are only a finite number of extreme optimal actions.

If $v, w \in \mathbb{R}^N$, by $v > w$ we shall mean: $v_k \geqslant w_k$ for all $k \in \{1, \ldots, N\}$ and $v_l > w_l$ for at least one $l \in \{1, \ldots, N\}$.

By 0_N we denote an N-vector with each component equal to 0 and by 0_{NN} $(N \times N)$-matrix is meant where each element equals 0.

3. Preliminaries

In this section some well-known facts will be recalled. Furthermore, two LP-problems and their duals will be stated, which form the body of our algorithm.

Lemma 3.1. *For stationary strategies $\tilde{\sigma}$ and $\tilde{\rho}$ we have*

$$\inf_{\pi_2} V(\tilde{\sigma}, \pi_2) = \min_{\rho} V(\tilde{\sigma}, \rho) = \min_{\rho^p} V(\tilde{\sigma}, \rho^p),$$

$$\sup_{\pi_1} V(\pi_1, \tilde{\rho}) = \max_{\sigma} V(\sigma, \tilde{\rho}) = \max_{\sigma^p} V(\sigma^p, \tilde{\rho}).$$

[Here σ^p and ρ^p are pure stationary strategies, i.e. $\sigma_k^p(i) \in \{0, 1\}$ and $\rho_k^p(j) \in \{0, 1\}$ for all $k \in S$, $i \in A_{1k}$, $j \in A_{2k}$.]

A proof of this lemma, which runs along the same lines as the proof of Theorem 1, p. 91 in Derman [2], can be found in Hordijk et al. [10].

We will now state the linear programming problem, which corresponds to a player 2 control stochastic game $\langle S, \{A_{1k}; k \in S\}, \{A_{2k}; k \in S\}, r, p\rangle$ (where $S_1 = \emptyset, S_2 = S$).

LP1. Variables: $g = (g_1, \ldots, g_N)$, $v = (v_1, \ldots, v_N)$, $x = \{x_k(i); k \in S, i \in A_{1k}\}$.

Max $\sum_{k \in S} g_k$ subject to

(i) $g_k - \sum_{l \in S} p(l|k, j) g_l \leqslant 0$ for all $k \in S$, $j \in A_{2k}$,

(ii) $g_k + v_k - \sum_{i \in A_{1k}} r(k, i, j) x_k(i) - \sum_{l \in S} p(l|k, j) v_l \leqslant 0$

for all $k \in S$, $j \in A_{2k}$

(iii) $\sum_{i \in A_{1k}} x_k(i) = 1$ for all $k \in S$,

(iv) $x_k(i) \geqslant 0$ for all $k \in S$, $i \in A_{1k}$.

DLP1. Dual variables: $d = (d_1, \ldots, d_N)$, $y = \{y_k(j); k \in S, j \in A_{2k}\}$, $z = \{z_k(j); k \in S, j \in A_{2k}\}$.

Min $\sum_{k \in S} d_k$ subject to

(j) $\sum_{k \in S} \sum_{j \in A_{2k}} (\delta_{kl} - p(l|k, j)) y_k(j) +$

$+ \sum_{j \in A_{2l}} z_l(j) = 1$ for all $l \in S$,

(jj) $\sum_{k \in S} \sum_{j \in A_{2k}} (\delta_{kl} - p(l|k, j)) z_k(j) = 0$ for all $l \in S$,

(jjj) $-\sum_{j \in A_{2k}} r(k, i, j) z_k(j) + d_k \geqslant 0$

for all $k \in S$, $i \in A_{1k}$,

(jv) $y_k(j), z_k(j) \geqslant 0$ for all $k \in S$, $j \in A_{2k}$.

[Here $\delta_{kl} := 1$ if $k = l$ and $\delta_{kl} := 0$ otherwise.]

Note that with an $x = (x_1, \ldots, x_N)$ obeying (iii) and (iv), one can associate a stationary strategy $\sigma_x = (\sigma_{x1}, \ldots, \sigma_{xN})$ for player 1 in an obvious way.

As shown in Vrieze [14], solving LP1 and DLP1 corresponds to solving the player 2 control stochastic game, with respect to the average payoff criterion. If $\hat{g}, \hat{v}, \hat{x}$ is a solution of LP1, then $\hat{g}$ is the value of the stochastic game and $\sigma_{\hat{x}}$ is an optimal strategy for player 1. Optimal stationary strategies for player 2 correspond to solutions of DLP1.

In the following, for a player 2 control stochastic game Γ, $R(\Gamma)$ will denote the sets of states k, for wich player 2 has an optimal stationary strategy ρ, such that state k is recurrent for $P(\rho)$.

Lemma 3.2. *Let $g = (g_1, g_2, \ldots, g_N)$ be the value for a player 2 control stochastic game. Then*

$$g_k = \min_{j \in A_{2k}} \sum_{l \in S} p(l|k, j) g_l \quad \textit{for all } k \in S.$$

Let $O_{2k} := \{j \in A_{2k}; g_k = \sum_{l \in S} p(l|k, j) g_l\}$. *Let* $v \in \mathbb{R}^N$ *be such that*

$$g_k + v_k \leqslant \operatorname{val}_{A_{1k} \times O_{2k}} [r(k, i, j) + \sum_{l \in S} p(l|k, j) v_l]$$

for all $k \in S$. (*)

Then the equality sign in (*) *holds for each* $k \in R(\Gamma)$.

Proof. The value $g = (g_1, g_2, \ldots, g_N)$ satisfies condition (i) of LP1. So $g_k \leqslant \min_{j \in A_{2k}} \sum_{l \in S} p(l|k, j) g_l$ for each $k \in S$. However in Vrieze [14], Lemma 2.9 it is shown that the equality sign holds for each $k \in S$, when g is part of an optimal solution to LP1.

Concerning the second assertion, take $\hat{k} \in R(\Gamma)$ and let ρ^* be an optimal stationary strategy for player 2 such that state $\hat{k}$ is recurrent with respect to $P(\rho^*)$. Let $E(\rho^*)$ be the ergodic set to which $\hat{k}$ belongs. By the first part of this theorem we have $g \leqslant P(\rho^*) \cdot g$. By multiplying both sides of this vector inequality with $Q(\rho^*)$ and remembering that $Q \cdot P = Q$, it follows that the equality sign holds at least in the components corresponding to the states belonging to $E(\rho^*)$. We denote the parts of g, P, Q, r and v corresponding to the set $E(\rho^*)$ by $\hat{g}$, $\hat{P}$, $\hat{Q}$, $\hat{r}$ and $\hat{v}$ respectively.

Then $\hat{g} = \hat{P}(\rho^*) \cdot \hat{g}$, hence $\hat{g} = \hat{P}^t(\rho^*) \cdot \hat{g}$ for each $t \in \mathbb{N}$ and then $\hat{g} = \hat{Q}(\rho^*) \cdot \hat{g}$.

Now suppose that the inequality sign in $(*)$ is strict for state $\hat{k}$. Then there exists a stationary strategy σ for player 1 such that

$$\hat{g} + \hat{v} < \hat{r}_{\sigma\rho^*} + \hat{P}(\rho^*) \cdot \hat{v}.$$

Multiplying this vector inequality by $\hat{Q}(\rho^*)$ yields:

$$\hat{g} = \hat{Q}(\rho^*) \cdot \hat{g} < \hat{Q}(\rho^*) \cdot \hat{r}_{\sigma\rho^*}.$$

Hence ρ^* cannot be optimal, which is a contradiction. This shows that for each $k \in R(\Gamma)$ the equality sign holds in $(*)$. □

With a player 2 control stochastic game we associate also another linear programming problem LP2. Because we will use this program for games, with payoffs of the type $r(k, i, j) - g_k$, it is convenient to incorporate this special form already at this place. So $g_1, \ldots, g_N$ in LP2 are not variables like in LP1 but constants.

LP2. Variables: $u = (u_1, u_2, \ldots, u_N)$, $x = \{x_k(i); k \in S, i \in A_{1k}\}$.

Max $\sum_{k \in S} u_k$ subject to

(i) $u_k - \sum_{i \in A_{1k}} (r(k, i, j) - g_k) x_k(i)$

$$- \sum_{l \in S} p(l|k, j) u_l \leqslant 0 \quad \text{for all } k \in S \text{ and } j \in A_{2k},$$

(ii) $\sum_{i \in A_{1k}} x_k(i) = 1$ and $x_k(i) \geqslant 0$

for all $k \in S$ and $i \in A_{1k}$.

DLP2. Dual variables: $b = (b_1, \ldots, b_N)$, $y = \{y_k(j); k \in S, j \in A_{2k}\}$.

Min $\sum_{k \in S} b_k$ subject to

(j) $\sum_{k \in S} \sum_{j \in A_{2k}} (\delta_{kl} - p(l|k, j)) y_k(j) = 1$

for all $l \in S$,

(jj) $- \sum_{j \in A_{2k}} (r(k, i, j) - g_k) y_k(j) + b_k \geqslant 0$

for all $k \in S$, $i \in A_{1k}$,

(jjj) $y_k(j) \geqslant 0$ for all $k \in S$, $j \in A_{2k}$.

Hordijk and Kallenberg [8] have shown that for the transient case (i.e. for the case where $\lim_{t \to \infty} P^t(\rho^p) = 0_{NN}$ for all pure stationary strategies ρ^p) LP2 is feasible and has a solution, which correspond to a solution of the game. We will need an extension of their results to what we call a semi-transient player 2 control stochastic game. That is a game with average payoff value 0_N, such that $\sum_{l \in S} p(l|k, j) \leqslant 1$ for all $j \in A_{2k}$ and all $k \in S$, and such that player 2 has a stationary strategy ρ, such that corresponding to $P(\rho)$ all states are transient.

Lemma 3.3. *For a semi-transient player 2 control stochastic game with payoffs of the form $r(k, i, j) - g_k$ and average payoff value 0_N the corresponding linear program LP2 is feasible and has a solution u^*, for which*

$$u_k^* = \operatorname{Val}_{A_{1k} \times A_{2k}} [r(k, i, j) - g_k + \sum_{l \in S} p(l|k, j) u_l^*]$$

for all $k \in S$.

Proof. Add a state $N + 1$, where both players have one action 1 and such that $p(N + 1 | N + 1, 1) = 1$, $r(N + 1, 1, 1) = 0$ and $p(N + 1 | k, j) = 1 - \sum_{l \in S} p(l|k, j)$ for all $k \in S$. Then we obtain a stochastic game with non-stopping transition probabilities and which obviously has also average payoff value 0_N. But this means (cf. Federgruen [3], theorem 7.4.1) that there exists a vector $v \in \mathbb{R}^{N+1}$, such that

$$v_k = \operatorname{Val}_{A_{1k} \times A_{2k}} [r(k, i, j) - g_k + \sum_{l \in S \cup \{N+1\}} p(l|k, j) v_l]$$

for all $k \in \{1, 2, \ldots, N + 1\}$.

Let $x_k = (x_k(1), x_k(2), \ldots, x_k(m_{1k}))$ be an optimal mixed action for the above matrix game. Then it can be easily checked that the pair (u, x) satisfies conditions (i) and (ii) of LP2 where $u_k = v_k - v_{N+1}$ for each $k \in S$. So LP2 is feasible. Next take an arbitrary feasible pair (u, x) for LP2. Let ρ be such that, corresponding to $P(\rho)$, all states are transient. Then condition (i) implies

$$u \leqslant r_{\sigma_x \rho} - g + P(\rho) u$$

and by iterating this inequality we obtain as a consequence of the transiency of all states:

$$u \leqslant \sum_{t=0}^{\infty} P^t(\rho) (r_{\sigma_x \rho} - g) \leqslant \sup_{\sigma} \sum_{t=0}^{\infty} P^t(\rho) (r_{\sigma\rho} - g).$$

Since $P(\rho)$ corresponds to a transient Markov chain, we may conclude that for the feasible solutions (u, x) of LP2, Σu_k is uniformly bounded. Now let (u^*, x^*) be an optimal solution of LP2 and suppose that there is a state k such that

$$u_k^* < \operatorname{Val}_{A_{1k} \times A_{2k}} [r(k, i, j) - g_k + \sum_{l \in S} p(l|k, j) u_l^*].$$

Let $\tilde{x}_k$ be an optimal action for player 1 for the matrix game in the right side of the above inequality. Then for $\epsilon > 0$, small enough, it follows that

$$u_k^* + \epsilon \leqslant \min_j \sum_{i \in A_{1k}} (r(k, i, j) - g_k) \tilde{x}_k(i) +$$

$$+ \sum_{l \in S \setminus \{k\}} p(l|k, j) u_l^* + p(k|k, j)(u_k^* + \epsilon).$$

But then the pair $(\bar{u}, \bar{x})$ with $\bar{u}_l = u_l^*$, $\bar{x}_l(i) = x_l^*(i)$ if $l \neq k$ and $\bar{u}_k = u_k + \epsilon$, $\bar{x}_k(i) = \tilde{x}_k(i)$ is feasible for LP2. Moreover $\sum_{l \in S} \bar{u}_l > \sum_{l \in S} u_l^*$. This leads to a contradiction since we have assumed, that (u^*, x^*) is an optimal solution of LP2. Hence the equality in the lemma holds. □

4. The Algorithm

In this section we will state a finite algorithm, which gives in a finite number of steps the solution for the switching control proplem $\Gamma = \langle S, S_1, S_2, \{A_{1k}; k \in S\}, \{A_{2k}; k \in S\}, r, p \rangle$.

The part of a stationary strategy σ of player 1, which refers to the set S_1, is denoted by σ^c. If we fix a particular σ^c, then the remaining game is a player 2 control stochastic game, denoted by $\tilde{\Gamma}(\sigma^c)$. Hence $\tilde{\Gamma}(\sigma^c) = \langle \tilde{S}, \{\tilde{A}_{1k}; k \in \tilde{S}\}, \{\tilde{A}_{2k}; k \in \tilde{S}\}, \tilde{r}, \tilde{p} \rangle$, where $\tilde{S} = S = S_1 \cup S_2$, where for $k \in S_1$: $\tilde{A}_{1k} = \{1\}$, $\tilde{A}_{2k} = A_{2k}$, $\tilde{r}(k, 1, j) = \sum_{i \in A_{1k}} r(k, i, j) \sigma_k^c(i)$, $\tilde{p}(l|k, j) = \sum_{i \in A_{1k}} p(l|k, i) \sigma_k^c(i)$, and for $k \in S_2$: $\tilde{A}_{1k} = A_{1k}$, $\tilde{A}_{2k} = A_{2k}$, $\tilde{r}(k, i, j) = r(k, i, j)$, $\tilde{p}(l|k, j) = p(l|k, j)$. The corresponding LP1-linear program for this game will be denoted by LP1 $(\tilde{\Gamma}(\sigma^c))$.

Now fix for a moment a subset $S_0 \subset S$, vectors g, $w \in \mathbb{R}^N$, a particular σ^c and for each $k \in S_0$ a non-empty subset O_{2k} of A_{2k}. Then corresponding to Γ and the five parameters S_0, g, w, σ^c and $\{O_{2k}; k \in S_0\}$ we introduce the player 2 control stochastic game $\bar{\Gamma}(S_0, g, w, \sigma^c, \{O_{2k}; k \in S_0\}) = \langle \bar{S}, \{\bar{A}_{1k}; k \in \bar{S}\}, \{\bar{A}_{2k}; k \in \bar{S}\}, \bar{r}, \bar{p} \rangle$ where $\bar{S} = S_0$, and where for $k \in \bar{S} \cap S_1$: $\bar{A}_{1k} = \{1\}$, $\bar{A}_{2k} = O_{2k}$, $\bar{r}(k, i, j) = -g_k + \sum_{i \in A_{1k}} (r(k, i, j) + \sum_{l \in S \setminus S_0} p(l|k, i) w_l) \sigma_k^c(i)$, $\bar{p}(l|k, j) = \sum_{i \in A_{1k}} p(l|k, i) \sigma_k^c(i)$ for $l \in S_0$, and where for $k \in \bar{S} \cap S_2$: $\bar{A}_{1k} = A_{1k}$, $\bar{A}_{2k} = O_{2k}$, $\bar{r}(k, i, j) = -g_k + r(k, i, j) + \sum_{l \in S \setminus S_0} p(l|k, j) w_l$ and $\bar{p}(l|k, j) = p(l|k, j)$ for $l \in S_0 = \bar{S}$.

The corresponding LP2-program of this game will be denoted be LP2$(\bar{\Gamma}(S_0, g, w, \sigma^c, \{O_{2k}; k \in S_0\}))$.

Now we have enough tools to establish our algorithm.

Algorithm

Step 1. Take $t = 0$ and choose $g(0) = (-M, \ldots, -M)$ (where $M = \max_{k,i,j} |r(k, i, j)|$), choose $w(0) = 0_N$, $S(0) = \emptyset$, $\sigma^c(0)$ such that for each $k \in S_1$ the action $\sigma_k^c(0)$ is an extreme optimal action for player 1 in the matrix game $[r(k, i, j)]_{A_{1k} \times A_{2k}}$.

Step 2. Take general t and the associated current values of the entities $g(t)$, $w(t)$, $S(t)$, $\sigma^c(t)$. Determine for each $k \in S_1$

$$O_{1k}(t+1) := \{i \in A_{1k}; \sum_{l \in S} p(l|k, i) g_l(t) =$$

$$= \max_{\tilde{i} \in A_{1k}} \sum_{l \in S} p(l|k, \tilde{i}) g_l(t)\}$$

and for each $k \in S_2$:

$$O_{2k}(t+1) := \{j \in A_{2k}; \sum_{l \in S} p(l|k, j) g_l(t) = g_k(t)\}.$$

Proceed to step 3.

Step 3. Choose $\sigma^c(t+1)$ such that for each $k \in S_1$, $\sigma_k^c(t+1)$ is an extreme optimal action for player 1 in the matrix game

$$\Lambda_{1k}(t) := [r(k, i, j) + \sum_{l \in S} p(l|k, i) w_l(t)]_{O_{1k}(t+1) \times A_{2k}}.$$

However, if $\operatorname{Car}(\sigma_k^c(t)) \subset O_{1k}(t+1)$ and if

$$g_k(t) + w_k(t) = \operatorname{Val}(\Lambda_{1k}(t)) = \min_j r(k, \sigma_k^c(t), j)$$

$$+ \sum_{l \in S} p(l|k, \sigma_k^c(t)) w_l(t)$$

then put $\sigma_k^c(t+1) := \sigma_k^c(t)$.

Step 4. Obtain $g(t+1)$, $v(t+1)$ by solving LP1 $(\tilde{\Gamma}(\sigma^c(t+1)))$.

Step 5. If $g(t+1) \neq g(t)$, then put $w(t+1) := v(t+1)$, $S(t+1) = \emptyset$ and return to step 2, taking $t := t+1$.

If $g(t+1) = g(t)$, then go to step 6.

Step 6. Put $O_{2k}(t+1) := A_{2k}$ for $k \in S_1$.

Let $G_1(t+1) := \{k \in S_1; g_k(t) + w_k(t) < \mathrm{Val}(\Lambda_{1k}(t))\}$

$G_2(t+1) := \{k \in S_2; g_k(t) + w_k(t) < \mathrm{Val}(\Lambda_{2k}(t))\}$

where $\Lambda_{2k}(t) = [r(k,i,j) + \sum_l p(l|k,j)w_l(t)]_{A_{1k} \times O_{2k}(t+1)}$.

Put $G(t+1) := G_1(t+1) \cup G_2(t+1)$.

If $G(t+1) = \emptyset$, then go to step 9. Else put $S(t+1) := G(t+1) \cup S(t)$ and go to step 7.

Step 7. Put $O_{2k}(t+1) := A_{2k}$ for $k \in S(t+1) \cap S_2$. Find $u_k(t+1)$ for each $k \in S(t+1)$ by solving for a semi-transient player 2 control stochastic game the LP problem LP2 $(\tilde{\Gamma}(S(t+1), g(t+1), w(t), \sigma^c(t+1), \{O_{2k}(t+1); k \in S(t+1)\}))$.

Step 8. Put $w_k(t+1) := w_k(t)$ if $k \notin S(t+1)$ and

$w_k(t+1) := u_k(t+1)$ if $k \in S(t+1)$.

Return to step 2 with $t := t+1$.

Step 9. The vector $g(t)$ is now the value vector for the original game. Further σ^* and ρ^* are optimal stationary strategies, if they are chosen as follows: for $k \in S_1$, σ_k^* and ρ_k^* must be optimal in the matrix game $\Lambda_{1k}(t)$, and for $k \in S_2$, σ_k^* and ρ_k^* must be optimal in the matrix game $\Lambda_{2k}(t)$.

In proving that in step 9 we indeed obtain a solution of the game we will show that in each stage $t = 0, 1, 2, \ldots$ the following eight properties are valid. Here $g(-1)$ is chosen, such that $g(-1) \leqslant g(0) = (-M, -M, \ldots, -M)$. We recall that $R(\Gamma)$, where Γ is a player 2 control stochastic game is defined as the set of states k for which player 2 has an optimal stationary strategy such that state k is recurrent with respect to $P(\rho)$.

$A_1(t)$: $g_k(t) \leqslant \sum_{l \in S} p(l|k, \sigma_k^c(t)) g_l(t)$ for each $k \in S_1$

$A_2(t)$: $g_k(t) \leqslant \sum_{l \in S} p(l|k,j) g_l(t)$ for each $k \in S_2, j \in A_{2k}$

$B_1(t)$: $g_k(t) + w_k(t) \leqslant r(k, \sigma_k^c(t), j) + \sum_{l \in S} p(l|k, \sigma_k^c(t)) w_l(t)$

for each $k \in S_1$ and $j \in A_{2k}$

$B_2(t)$: $g_k(t) + w_k(t) \leqslant \mathrm{Val}(\Lambda_{2k}(t))$ for each $k \in S_2$

$C(t)$: $g(t) \geqslant g(t-1)$

$D(t)$: If $g(t) = g(t-1)$, then $R(\tilde{\Gamma}(\sigma^c(t))) \subset R(\tilde{\Gamma}(\sigma^c(t-1)))$, and $\sigma_k^c(t) = \sigma_k^c(t-1)$ for each $k \in R(\tilde{\Gamma}(\sigma^c(t)))$

$E(t)$: $S(t) \cap R(\tilde{\Gamma}(\sigma^c(t))) = \emptyset$

$F(t)$: If $g(t) = g(t-1)$ and $G(t) \neq \emptyset$ then $w(t) > w(t-1)$.

Since $g(-1) \leqslant g(0)$, it follows that $A_1(0)$, $A_2(0)$, $B_1(0)$, $B_2(0)$, $C(0)$, $D(0)$, $E(0)$ and $F(0)$ hold. By induction on t we want to prove that $A_1(t), \ldots, F(t)$ hold for each $t \in \{0, 1, \ldots\}$. To this purpose, we need a string of lemmas.

Lemma 4.1. *Suppose* $g_k(t) = \max_{i \in A_{1k}} \sum_{l \in S} p(l|k,i) g_l(t)$ for $k \in S_1$. *Then* $\mathrm{Car}(\sigma_k^c(t)) \subset O_{1k}(t+1)$.

If, furthermore, property $B_1(t)$ *holds, then for all* $j \in A_{2k}$

$$g_k(t) + w_k(t) \leqslant r(k, \sigma_k^c(t+1), j) + \sum_{l \in S} p(l|k, \sigma_k^c(t+1)) w_l(t).$$

Proof. Condition (i) of LP1$(\tilde{\Gamma}(\sigma^c(t)))$ yields: $g_k(t) \leqslant \sum_{l \in S} p(l|k, \sigma_k^c(t)) g_l(t)$, which in combination with the assumption in the lemma can only be true if $g_k(t) = \sum_{l \in S} p(l|k, \tilde{i}) g_l(t)$ for each $\tilde{i} \in \mathrm{Car}(\sigma_k^c(t))$. Hence $\mathrm{Car}(\sigma_k^c(t) \subset \subset O_{1k}(t+1)$. This fact in combination with $B_1(t)$ implies

$$g_k(t) + w_k(t) \leqslant \min_j (r(k, \sigma_k^c(t), j) + \sum_l p(l|k, \sigma_k^c(t)) w_l(t)) \leqslant \mathrm{Val}\, \Lambda_{1k}(t) =$$

$$= \min_j (r(k, \sigma_k^c(t+1), j) + \sum_{l \in S} p(l|k, \sigma_k^c(t+1) w_l(t)). \quad \square$$

Lemma 4.2. *Properties* $A_1(t+1)$ *and* $A_2(t+1)$ *hold.*

Proof. This is an immediate consequence of condition (i) of LP1$(\tilde{\Gamma}(\sigma^c(t+1)))$. $\square$

Lemma 4.3. *Suppose that* $A_1(t)$, $A_2(t)$, $B_1(t)$ *and* $B_2(t)$ *hold. Then* $C(t+1)$ *holds.*

Proof. Choose the stationary strategy $\tilde{\sigma} = (\tilde{\sigma}_1, \ldots, \tilde{\sigma}_N)$ as follows. If $k \in S_1$, then take $\tilde{\sigma}_k = \sigma_k^c(t+1)$, and if $k \in S_2$, then let $\tilde{\sigma}_k$ be an optimal action in the matrix game $\Lambda_{2k}(t)$. Let ρ^p be an arbitrary pure stationary strategy. It is sufficient to show that $V(\tilde{\sigma}, \rho^p) \geqslant g(t)$. By $A_1(t)$, $A_2(t)$ and step 2 of the algorithm, we have

$$g(t) \leqslant P_{\tilde{\sigma}\rho p}\, g(t). \qquad (1)$$

A consequence of (1) is that the equality sign holds for the coordinates corresponding to the recurrent states of $P_{\tilde{\sigma}\rho p}$. We denote this set of states by $R(\tilde{\sigma}, \rho^p)$. For $k \in R(\tilde{\sigma}, \rho^p) \cap S_1$, this yields:

$$g_k(t) = \sum_{l \in S} p(l|k, \sigma_k^c(t+1)) g_l(t) =$$

$$= \max_i \sum_{l \in S} p(l|k, i) g_l(t).$$

So we may apply Lemma 4.1, obtaining

$$g_k(t) + w_k(t) \leqslant r(k, \tilde{\sigma}_k, \rho_k^p) + \sum_{l \in S} p(l|k, \tilde{\sigma}_k) w_l(t). \quad (2)$$

For $k \in R(\tilde{\sigma}, \rho^p) \cap S_2$ we have by $B_2(t)$ and by the choice of $\tilde{\sigma}$, noting that $\mathrm{Car}(\rho_k^p) \subset O_{2k}(t+1)$:

$$g_k(t) + w_k(t) \leqslant r(k, \tilde{\sigma}_k, \rho_k^p) + \sum_{l \in S} p(l|k, \rho_k^p) w_l(t). \quad (3)$$

The inequalities (1), (2) and (3) imply: $g(t) \leqslant Q_{\tilde{\sigma}\rho p} r_{\tilde{\sigma}\rho p} = V(\tilde{\sigma}, \rho^p)$. Hence $g(t+1) \geqslant g(t)$. □

Lemma 4.4. *Suppose $A_1(t)$, $B_1(t)$ and $B_2(t)$ hold. Then $D(t+1)$ holds.*

Proof. Suppose $g(t+1) = g(t)$. Observe that in $\tilde{\Gamma}(\sigma^c(t+1))$ player 2 has in the states belonging to S_1 no influence on the transitions. Then by Lemma 3.2 we have $g_k(t) = \sum_{l \in S} p(l|k, \sigma_k^c(t+1) \cdot g_l(t)$ for each $k \in S_1$. Since $\mathrm{Car}(\sigma_k^c(t+1)) \subset O_{1k}(t+1)$ this implies $g_k(t) = \max_{i \in A_{1k}} \sum_{l \in S} p(l|k, i) \cdot g_l(t)$ for all $k \in S_1$. Hence, by Lemma 4.1 for each $k \in S_1$:

$$g_k(t) + w_k(t) \leqslant \min_j r(k, \sigma_k^c(t+1), j) + \sum_{l \in S} p(l|k, \sigma_k^c(t+1)) w_l(t). \quad (4)$$

Since $g(t+1) = g(t)$ equals the value of $\tilde{\Gamma}(\sigma^c(t+1))$, Lemma 3.2 can be applied to (4) and $B_2(t)$, implying that for $k \in R(\tilde{\Gamma}(\sigma^c(t+1)))$ the equality sign holds in the respective inequalities, which shows that for $k \in R(\tilde{\Gamma}(\sigma^c(t+1)))$ the action $\sigma_k^c(t)$ is optimal in the matrix game $\Lambda_{1k}(t)$. So by step 3 of the algorithm

$$\sigma_k^c(t+1) = \sigma_k^c(t) \quad \text{for all } k \in R(\tilde{\Gamma}(\sigma^c(t+1))) \cap S_1. \quad (5)$$

Fix $k \in R(\tilde{\Gamma}(\sigma^c(t+1)))$ and let ρ be optimal for player 2 in $\tilde{\Gamma}(\sigma^c(t+1))$ and such that k is a recurrent state with respect to ρ. Then (5) implies, that in the ergodic set to which k belongs, ρ is also optimal in $\tilde{\Gamma}(\sigma^c(t))$, and clearly the state k is recurrent, which shows that $R(\tilde{\Gamma}(\sigma^c(t+1))) \subset R(\tilde{\Gamma}(\sigma^c(t)))$. □

Lemma 4.5. *Suppose $A_1(t)$, $B_1(t)$, $B_2(t)$ and $E(t)$ hold. Then $E(t+1)$ holds.*

Proof. If $g(t+1) > g(t)$, then $S(t+1) = \emptyset$, and then $E(t+1)$ is true. Hence, suppose $g(t+1) = g(t)$. From $E(t)$ and $R(\tilde{\Gamma}(\sigma^c(t+1))) \subset R(\tilde{\Gamma}(\sigma^c(t)))$ (Lemma 4.4) it follows that $S(t) \cap R(\tilde{\Gamma}(\sigma^c(t+1))) = \emptyset$. It now suffices to show that $G(t+1) \cap R(\tilde{\Gamma}(\sigma^c(t+1))) = \emptyset$. Using Lemma 4.1 and Lemma 3.2 we obtain:

$$g_k(t) + w_k(t) = \mathrm{Val}(\Lambda_{1k}(t)) \quad \text{for } k \in R(\tilde{\Gamma}(\sigma^c(t+1))) \cap S_1$$

and

$$g_k(t) + w_k(t) = \mathrm{Val}(\Lambda_{2k}(t)) \quad \text{for } k \in R(\tilde{\Gamma}(\sigma^c(t+1))) \cap S_2.$$

Hence it follows from the definition of $G(t+1)$ in step 6 that $G(t+1) \cap R(\tilde{\Gamma}(\sigma^c(t+1))) = \emptyset$. □

Lemma 4.6. *Suppose $A_1(t)$, $B_1(t)$, $B_2(t)$ and $E(t)$ hold. Then $F(t+1)$ holds.*

Proof. Suppose $g(t+1) = g(t)$ and $G(t+1) \neq \emptyset$. From $A_1(t)$, $B_1(t)$ and Lemma 4.1 we obtain that for $k \in S(t+1) \cap S_1$:

$$g_k(t) + w_k(t) \leqslant \min_j (r(k, \sigma_k^c(t+1), j) + \sum_{l \in S} p(l|k, \sigma_k^c(t+1)) w_l(t) \quad (6)$$

and for $k \in S(t+1) \cap S_2$ we see from $B_2(t)$:

$$g_k(t) + w_k(t) \leqslant \mathrm{Val}(\Lambda_{2k}(t)). \quad (7)$$

As $G(t+1) \neq \emptyset$, the strict inequality holds in (6) and (7) at least in one component.

Since the value of $\tilde{\Gamma}(\sigma^c(t+1))$ equals $g(t+1) = g(t)$ and since $S(t+1) \cap R(\tilde{\Gamma}(\sigma^c(t+1))) = \emptyset$ (Lemma 4.5) it can be verified that the game $\tilde{\Gamma}(S(t+1), g(t+1), w(t), \sigma^c(t+1), \{O_{2k}(t+1); k \in S(t+1)\})$ is a semi-transient player 2 control stochastic game. Namely

(a) Obviously $\sum_{l \in S(t+1)} \bar{p}(l|k, j) \leqslant 1$ for $k \in S(t+1)$,

(b) the part corresponding to $S(t+1)$ of an optimal stationary strategy of player 2 in the game $\tilde{\Gamma}(\sigma^c(t+1))$, gives when applied in $\bar{\Gamma}(.,.,.,.,.)$ a transient stochastic matrix and

(c) the average reward value equals 0_N. By (b) it follows that the value is at most 0_N. If a stationary strategy ρ for player 2 in $\bar{\Gamma}(.,.,.,.,.)$ is such that some states of $S(t+1)$ are recurrent, then ρ is bad for player 2 in view of $S(t+1) \cap R(\tilde{\Gamma}(\sigma^c(t+1))) = \emptyset$. Hence the best player 2 can do in $\bar{\Gamma}$ is playing a transient stationary strategy, resulting in value 0_N.

Let for $k \in S(t+1) \cap S_2$, $\hat{x}_k$ be an optimal action for player 1 in $\Lambda_{2k}(t)$. Then, putting $\hat{x}_k(1) = 1$ if $k \in S(t+1) \cap S_1$, it can be seen that the pair $(\{w_k(t); k \in S(t+1)\}, \{\hat{x}_k(i);\ k \in S(t+1),\ i \in \bar{A}_{1k}\})$ satisfies conditions (i) and (ii) of LP2$(\bar{\Gamma}(.,.,.,.,.))$. But in (6) and (7) at least one strict inequality sign holds. Hence we obtain for the solution $\{u_k; k \in S(t+1)\}$ of this LP2 problem: $u_k \geqslant w_k(t)$ for all $k \in S(t+1)$ with the inequality sign holding for at least one coordinate (cf. Lemma 3.3). □

Lemma 4.7. *Suppose $A_1(t)$, $B_1(t)$, $B_2(t)$ and $E(t)$ holds. Then $B_1(t+1)$ and $B_2(t+1)$ hold.*

Proof. If $g(t+1) \neq g(t)$, then $B_1(t+1)$ and $B_2(t+1)$ follow from the condition (ii) of LP1$(\tilde{\Gamma}(\sigma^c(t+1)))$. Suppose now $g(t+1) = g(t)$. From $F(t+1)$ (Lemma 4.6) we get $w_k(t+1) \geqslant w_k(t)$ for each $k \in S(t+1)$. By definition $w_k(t+1) = w_k(t)$ for each $k \in S \setminus S(t+1)$. So using $B_1(t)$, $B_2(t)$ and Lemma 4.1 we then have that $B_1(t+1)$ and $B_2(t+1)$ hold for $k \in S \setminus S(t+1)$. However by condition (i) of LP2$(\bar{\Gamma}(S(t+1), g(t+1), w(t), \sigma^c(t+1), \{O_{2k}(t+1); k \in S(t+1)\}))$ it follows, that $B_1(t+1)$ and $B_2(t+1)$ also hold for $k \in S(t+1)$. □

Now, combining the Lemmas 4.1–4.7, we may conclude that the assumption "$A_1(t)$, $A_2(t)$, $B_1(t)$, $B_2(t)$, $C(t)$, $D(t)$, $E(t)$ and $F(t)$ hold" follows that "$A_1(t+1)$, $A_2(t+1)$, $B_1(t+1)$, $B_2(t+1)$, $C(t+1)$, $D(t+1)$, $E(t+1)$ and $F(t+1)$ hold". Hence we have

Theorem 4.8. *For each $t \in \{0, 1, 2, \ldots\}$ the properties $A_1(t), A_2(t), B_1(t), B_2(t), C(t), D(t), E(t)$ and $F(t)$ hold.*

Important is the following

Theorem 4.9. *The algorithm stops after a finite number of iterations.*

Proof. Parthasarathy and Raghavan [11] have shown that an extreme optimal action for player 1 in a matrix game of payoff type $[f(i,j) + h(i)]_{A \times B}$ is also an extreme optimal action for player 1 in some subgame $[f(i,j)]_{\alpha \times B}$ with $\alpha \subset A$ (cf. [11], Lemma 4.1, p. 381). Applied to step 3 of our algorithm, this means that for each state $k \in S_1$ an extreme optimal action will be chosen of some matrix game $[r(k,i,j)]_{\alpha_{1k}(t) \times A_{2k}}$, where $\alpha_{1k}(t) \subset A_{1k}$.

Shapley and Snow have shown that a matrix game has only a finite number of extreme optimal actions. Furthermore, a matrix game has a finite number of submatrices and there are a finite number of states, which ensures for each t that

the set from which $\sigma^c(t)$ is chosen is a finite one. (8)

It remains to show, that no cycles can occur, i.e. that no strategy repeats infinitely often.

By the properties $C(t)$ and $F(t)$ we can see that for each t exactly one of the following events occurs:

H1: $g(t) > g(t-1)$,

H2: $g(t) = g(t-1)$, $\sigma^c(t) \neq \sigma^c(t-1)$, $G(t) \neq \emptyset$, $w(t) > w(t-1)$,

H3: $g(t) = g(t-1)$, $\sigma^c(t) = \sigma^c(t-1)$, $G(t) \neq \emptyset$, $w(t) > w(t-1)$,

H4: $g(t) = g(t-1)$, $\sigma^c(t) = \sigma^c(t-1)$, $G(t) = \emptyset$.

Since $\tilde{\Gamma}(\sigma^c(t))$ only depends on $\sigma^c(t)$ we have in view of $C(t)$:

if H1 occurs on t, then $\sigma^c(m) \neq \sigma^c(n)$, (9)
$m \in \{t, t+1, \ldots\}$ and $n \in \{t-1, t-2, \ldots, 0\}$

Now suppose that from stage t, H2 repeats itself infinitely often. Since $|S(t)| \leqslant z - 1$ we may assume without loss of generality that $S(t) = S(t+1) = S(t+2) = \ldots$ But then observe that the optimal value of

LP2$(\bar{\Gamma}(S(t-1+n), g(t-1+n), w(t-2+n), \sigma^c(t-1+n), \{O_{2k}(t-1+n); k \in S(t-1+n)\}))$ in step 7 of the algorithm only depends on $\{\sigma_k^c(t-1+n); k \in S(t-1+n) = S(t)\}$, for each $n = 1, 2, \ldots$ for the other parameters do not change.

But since $w(t-1+l) > w(t-2+l)$ we get $\sigma^c(m+n) \neq \sigma^c(m)$, for $n = 1, 2, \ldots$ and $m = t-1, t, t+1, \ldots$ But then in view of (9) we see:

H2 can not repeat itself infinitely often (10)

Let n be the first time H2 does not occur. Then either $S(n) = \emptyset$ in which case H1 occurs, or it happens that H4 occurs, or possibly H3 occurs.

If H3 occurs on t, then by the construction of $G_1(t)$ and $G_2(t)$ and by the equality in Lemma 3.3 we see that $G(t) \cap S(t-1) = \emptyset$. Hence:

If H3 occurs then S(t) strictly includes $S(t-1)$ (11)

As last statement we have:

If H4 occurs then the algorithm stops (12)

Now observe that from (10) and (11) we may conclude that a sequence in which only the events H2 and H3 occur can not happen. But then in view of (9) it lasts a finite number of iterations before H4 occurs, which by (12) proves the theorem. □

Theorem 4.10. *Step 9 of the algorithm is reached after a finite number of iterations and provides a solution to the game, i.e. $g(t)$ equals the value of the game and ρ^* and σ^* are optimal stationary strategies for player 1 and player 2 respectively.*

Proof. By Theorem 4.9 step 9 is reached in a finite number of iterations. From $g(t+1) = g(t)$ it follows (cf. the proof of Lemma 4.4):

$$g_k(t) = \max_{i \in A_{1k}} \sum_{l \in S} p(l|k, i) \cdot g_l(t) \quad \text{for each } k \in S_1 \quad (13)$$

Next observe from Lemma 3.2 that

$$g_k(t) = \min_{j \in A_{2k}} \sum_{l \in S} p(l|k, j) \cdot g_l(t) \quad \text{for each } k \in S_2. \quad (14)$$

From the definitions of σ^* and ρ^* we know:

$$\mathrm{Car}(\sigma_k^*) \subset O_{1k}(t+1), \quad k \in S_1 \quad \text{and}$$
$$\mathrm{Car}(\rho_k^*) \subset O_{2k}(t+1), \quad k \in S_2. \quad (15)$$

From Lemma 4.1, property $B_2(t)$ and $G(t+1) = \emptyset$ we derive

$$g_k(t) + w_k(t) = \mathrm{Val}(\Lambda_{1k}(t)), \quad k \in S_1 \quad (16)$$

and

$$g_k(t) + w_k(t) = \mathrm{Val}(\Lambda_{2k}(t)), \quad k \in S_2. \quad (17)$$

Let ρ^p be an arbitrary pure stationary strategy of player 2. Then from (13)–(17) and the definition of σ^* we infer:

$$g(t) \leqslant P_{\sigma^* \rho^p} g(t) \quad (18)$$

$$g_k(t) + w_k(t) \leqslant r(k, \sigma_k^*, \rho_k^p) + \sum_{l \in S} p(l|k, \sigma_k^*) w_l(t), \quad k \in S_1 \quad (19)$$

$$g_k(t) + w_k(t) \leqslant r(k, \sigma_k^*, \rho_k^p) + \sum_{l \in S} p(l|k, \rho_k^p) w_l(t), \quad k \in S_2. \quad (20)$$

Similarly for an arbitrary pure stationary strategy σ^p of player 1 we derive from (13)–(17):

$$g(t) \geqslant P_{\sigma^p \rho^*} g(t) \quad (21)$$

$$g_k(t) + w_k(t) \geqslant r(k, \sigma_k^p, \rho_k^*) + \sum_{l \in S} p(l|k, \rho_k^p) w_l(t), \quad k \in S_1) \quad (22)$$

$$g_k(t) + w_k(t) \geqslant r(k, \sigma_k^p, \rho_k^*) + \sum_{l \in S} p(l|k, \rho_k^*) w_l(t), \quad k \in S_2. \quad (23)$$

Now (18)–(23) imply

$$Q_{\sigma^* \rho p}\, r_{\sigma^* \rho p} \geqslant g(t) \geqslant Q_{\sigma p \rho^*}\, r_{\sigma p \rho^*}. \quad (24)$$

Since σ^p and ρ^p are arbitrary, application of Lemma 3.1 to (24) results in

$$\min_{\pi_2} V(\sigma^*, \pi_2) \geqslant g(t) \geqslant \max_{\pi_1} V(\pi_1, \rho^*),$$

which shows the theorem. □

We conclude this paper with the remark, that our algorithm provides a constructive proof of the existence of the value and of optimal stationary strategies for both players for the switching control stochastic game. Also the fact, proved by Filar [4], [5], that player 1 (2) has an optimal stationary strategy σ^* (ρ^*), such that for $k \in S_1$ ($k \in S_2$) σ_k^* (ρ_k^*) is an optimal action in a matrix game of the form $[r(k, i, j)]_{\alpha_{1k} \times A_{2k}}$ ($[r(k, i, j)]_{A_{1k} \times \alpha_{2k}}$), where $\alpha_{1k} \subset A_{1k}$ ($\alpha_{2k} \subset A_{2k}$)), can be derived from our algorithm. Furthermore the finiteness of the algorithm proves the ordered field property (cf. [5]).

References

1. Bewley T, Kohlberg E (1978) On stochastic games with stationary optimal strategies. Math Oper Res 3:104–125
2. Derman C (1970) Finite state Markovian decision processes. Academic Press, New York
3. Federgruen A (1978) Markovian control problems. Ph.D. Dissertation, Math. Centre, Amsterdam
4. Filar JA (1979) Algorithms for solving some undiscounted stochastic games. Ph.D. Dissertation, University of Illinois, Chicago
5. Filar JA (1981) Ordered field property for stochastic games when the player who controls transitions changes from state to state. JOTA 34:503–515
6. Filar JA, Raghavan TES (1979) An algorithm for solving an undiscounted stochastic game in which one player controls transitions. Research Memorandum, University of Illinois, Chicago
7. Filar JA, Raghavan TES (1980) Two remarks concerning